(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) Übersetzung der **europäischen Patentschrift**

(97) **EP 0 625 775 B1**

(10) **DE 694 25 776 T 2**

(51) Int. Cl.[7]: **G 10 L 15/20**

(21) Deutsches Aktenzeichen: 694 25 776.1
(96) Europäisches Aktenzeichen: 94 104 846.4
(96) Europäischer Anmeldetag: 28. 3. 1994
(97) Erstveröffentlichung durch das EPA: 23. 11. 1994
(97) Veröffentlichungstag der Patenterteilung beim EPA: 6. 9. 2000
(47) Veröffentlichungstag im Patentblatt: 12. 4. 2001

(30) Unionspriorität:
62972 18. 05. 1993 US

(73) Patentinhaber:
International Business Machines Corp., Armonk, N.Y., US

(74) Vertreter:
Teufel, F., Dipl.-Phys., Pat.-Anw., 70569 Stuttgart

(84) Benannte Vertragstaaten:
DE, FR, GB

(72) Erfinder:
Epstein, Edward A., Putnam Valley, New York 10579, US

(54) Spracherkennungseinrichtung mit verbesserter Ausschliessung von Wörtern und Tönen welche nicht im Vokabular enthalten sind

DE 694 25 776 T 2

B E S C H R E I B U N G

Spracherkennungseinrichtung mit verbesserter Ausschließung von Wörtern und Tönen, die nicht im Vokabular enthalten sind

Grundlagen der Erfindung

Die Erfindung betrifft die Computerspracherkennung, insbesondere die Erkennung gesprochener Computerbefehle. Wenn ein gesprochener Befehl erkannt wird, führt der Computer eine oder mehrere dem Befehl zugeordnete Funktionen aus.

Im Allgemeinen besteht eine Spracherkennungsvorrichtung aus einem Akustikprozessor und einem gespeicherten Satz akustischer Modelle. Der Akustikprozessor misst Tonmerkmale einer Äußerung. Jedes akustische Modell stellt die akustischen Merkmale einer Äußerung eines oder mehrerer dem Modell zugeordneter Worte dar. Die Tonmerkmale der Äußerung werden mit jedem akustischen Modell verglichen, um einen Vergleichswert zu erzeugen. Der Vergleichswert für eine Äußerung und ein akustisches Modell ist eine Schätzung der Genauigkeit der Tonmerkmale der Äußerung im Vergleich zum akustischen Modell.

Das Wort bzw. die Worte, die dem akustischen Modell mit dem besten Vergleichswert zugeordnet werden, können als Erkennungsergebnis ausgewählt werden. Alternativ kann der

- akustische Vergleichswert mit anderen Vergleichswerten kombiniert werden, beispielsweise mit zusätzlichen akustischen Vergleichswerten und Sprachmodellvergleichswerten. Das Wort bzw. die Worte, die dem(den) akustischen Modell(en) mit dem

besten kombinierten Vergleichswert zugeordnet werden, können als Erkennungsergebnis ausgewählt werden.

Bei Befehls- und Steueranwendungen erkennt die Spracherkennungsvorrichtung vorzugsweise einen geäußerten Befehl, und das Computersystem führt den Befehl anschließend sofort aus, um eine dem erkannten Befehl zugeordnete Funktion auszuführen. Zu diesem Zweck kann der Befehl, der dem akustischen Modell mit dem besten Vergleichswert zugeordnet wird, als Erkennungsergebnis ausgewählt werden.

Ein schwerwiegendes Problem bei solchen Systemen besteht jedoch darin, dass unbeabsichtigte Töne, beispielsweise Husten, Seufzer oder gesprochene Worte, die nicht zur Erkennung vorgesehen sind, fälschlicherweise als gültige Befehle erkannt werden. Das Computersystem führt die falsch erkannten Befehle sodann sofort aus, um die zugeordneten Funktionen mit unbeabsichtigten Folgen auszuführen.

US-A-4 239 936 beschreibt ein Spracherkennungssystem, in dem die Intensität von Umgebungsgeräusch parallel zu den eingegebenen Sprachsignalen gemessen wird, wobei jedes dem eingegebenen Sprachsignal zugeordnetes Erkennungsergebnis zurückgewiesen wird, wenn die Intensität des Geräusches einen festgelegten Standardwert überschreitet.

Zusammenfassung der Erfindung

Eine Aufgabe der Erfindung ist die Bereitstellung einer Vorrichtung und eines Verfahrens zur Spracherkennung, das eine hohe Wahrscheinlichkeit aufweist, akustische Übereinstimmungen mit unbeabsichtigten Tönen oder gesprochenen Worten, die nicht

für die Spracherkennungseinrichtung vorgesehen sind, auszuschließen.

Eine andere Aufgabe der Erfindung ist die Bereitstellung einer Vorrichtung und eines Verfahrens zur Spracherkennung, das das akustische Modell kennzeichnet, das am besten mit einem Ton übereinstimmt und das eine hohe Wahrscheinlichkeit hat, das am besten übereinstimmende akustische Modell auszuschließen, falls der Ton unbeabsichtigt oder nicht für die Spracherkennungseinrichtung vorgesehen ist, das jedoch eine hohe Wahrscheinlichkeit hat, das am besten übereinstimmende akustische Modell anzunehmen, falls der Ton ein oder mehrere zur Erkennung vorgesehene Worte darstellt.

Eine Spracherkennungsvorrichtung gemäß der Erfindung umfasst einen Akustikprozessor zum Messen des Wertes von mindestens einem Merkmal von jeder aus einer Folge von mindestens zwei Tönen. Der Akustikprozessor misst den Wert des Merkmals von jedem Ton während jeder aus einer Reihe aufeinanderfolgender Zeitintervalle, um eine Folge von Merkmalsignalen zu erzeugen, die die Merkmalwerte des Tons darstellen. Außerdem werden Mittel zur Speicherung eines Satzes akustischer Merkmale bereitgestellt. Jedes akustische Befehlsmodell stellt eine oder mehrere Folgen akustischer Merkmalwerte dar, die eine Äußerung eines dem akustischen Befehlsmodell zugeordneten Befehls darstellen.

Ein Vergleichswertprozessor erzeugt einen Vergleichswert für jeden Ton und jedes von einem oder mehreren Befehlsmodellen aus dem Satz akustischer Befehlsmodelle. Jeder Vergleichswert umfasst eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Befehlsmodell und einer Reihe dem Ton

entsprechender Merkmalsignale. Es werden Mittel zum Ausgeben eines Erkennungssignals bereitgestellt, das dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entspricht, falls der beste Vergleichswert für den aktuellen Ton besser als ein Erkennungsschwellenwert für den aktuellen Ton ist. Die Erkennungsschwelle für den aktuellen Ton umfasst (a) einen ersten Vertrauenswert, falls der beste Vergleichswert für einen früheren Ton besser als eine Erkennungsschwelle für diesen früheren Ton war, oder (b) einen zweiten Vertrauenswert, der besser als der erste Vertrauenswert ist, falls der beste Vergleichswert für einen früheren Ton schlechter als die Erkennungsschwelle für diesen früheren Ton war.

Vorzugsweise tritt der frühere Ton unmittelbar vor dem aktuellen Ton auf.

Eine Spracherkennungsvorrichtung gemäß der Erfindung kann außerdem Mittel zur Speicherung von mindestens einem akustischen Schweigemodell umfassen, das eine oder mehrere Folgen akustischer Merkmalwerte darstellt, die das Fehlen einer gesprochenen Äußerung darstellen. Der Vergleichswertprozessor erzeugt außerdem einen Vergleichswert für jeden Ton und das akustische Schweigemodell. Jeder Schweigevergleichswert umfasst eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Schweigemodell und einer Reihe dem Ton entsprechender Merkmalsignale.

In diesem Aspekt der Erfindung umfasst die Erkennungsschwelle für den aktuellen Ton den ersten Vertrauenswert (a1), falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als eine Schweigevergleichsschwelle ist,

und falls der frühere Ton eine Dauer hat, die eine Schweigedauerschwelle überschreitet, oder (a2) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die geringer als die Schweigedauerschwelle ist, und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell besser als eine Erkennungsschwelle für diesen nächsten früheren Ton war, oder (a3) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist, und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell besser als eine Erkennungsschwelle für diesen früheren Ton war.

Die Erkennungsschwelle für den aktuellen Ton umfasst den zweiten Vertrauenswert, der besser als der erste Vertrauenswert ist, (b1) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist, und falls der frühere Ton eine Dauer hat, die geringer als die Schweigedauerschwelle ist, und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen nächsten früheren Ton war, oder (b2) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist, und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen früheren Ton war.

Das Erkennungssignal kann beispielsweise ein Befehlssignal zum Aufrufen eines dem Befehl zugeordneten Programms sein. In einem Aspekt der Erfindung umfasst das Ausgabemittel eine Anzeige, und das Ausgabemittel zeigt ein oder mehrere Worte an, die dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entsprechen, falls der beste Vergleichswert für den aktuellen Ton besser als der Erkennungsschwellenwert für den aktuellen Ton ist.

In einem anderen Aspekt der Erfindung gibt das Ausgabemittel ein Anzeigesignal für einen nicht erkennbaren Ton aus, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist. Das Ausgabemittel kann beispielsweise eine Anzeige für einen nicht erkennbaren Ton ausgeben, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist. Die Anzeige für einen nicht erkennbaren Ton kann beispielsweise ein oder mehrere Fragezeichen umfassen.

Der Akustikprozessor in der Spracherkennungsvorrichtung gemäß der Erfindung kann u.a. ein Mikrofon umfassen. Jeder Ton kann beispielsweise ein Vokalton sein, und jeder Befehl kann mindestens ein Wort umfassen.

Gemäß einem weiteren Aspekt der Erfindung wird ein Spracherkennungsverfahren bereitgestellt, wie es in Anspruch 11 definiert wird.

Gemäß der Erfindung können akustische Vergleichsprozessoren folglich in drei Kategorien unterteilt werden. Wenn der beste Vergleichswert besser als ein "guter" Vertrauenswert ist,

entspricht das Wort bzw. die Worte, die dem akustischen Modell mit dem besten Vergleichswert entsprechen, fast immer den gemessenen Tönen. Andererseits entspricht das Wort, das dem akustischen Modell mit dem besten Vergleichswert entspricht, fast nie den gemessenen Tönen, falls der beste Vergleichswert schlechter als ein "schlechter" Vertrauenswert ist. Wenn der beste Vergleichswert besser als der "schlechte" Vertrauenswert, jedoch schlechter als der "gute" Vertrauenswert ist, entspricht das Wort, das dem akustischen Modell mit dem besten Vergleichswert entspricht, mit hoher Wahrscheinlichkeit dem gemessenen Ton, wenn für das zuvor erkannte Wort angenommen wurde, da es eine hohe Wahrscheinlichkeit hat, dem vorhergehenden Ton zu entsprechen. Wenn der beste Vergleichswert besser als der "schlechte" Vertrauenswert, jedoch schlechter als der "gute" Vertrauenswert ist, entspricht das Wort, das dem akustischen Modell mit dem besten Vergleichswert entspricht, mit geringer Wahrscheinlichkeit dem gemessenen Ton, wenn das zuvor erkannte Wort ausgeschlossen wurde, da es eine geringe Wahrscheinlichkeit hat, dem vorhergehenden Ton zu entsprechen. Falls jedoch zwischen einem zuvor ausgeschlossenen Wort und dem aktuellen Wort mit dem besten Vergleichswert, der besser als der "schlechte" Vertrauenswert, jedoch schlechter als der "gute" Vertrauenswert ist, genügend Schweigen liegt, wird das aktuelle Wort ebenfalls mit einer hohen Wahrscheinlichkeit, dem gemessenen aktuellen Ton zu entsprechen, angenommen.

Durch die Annahme der Vertrauenswerte gemäß der Erfindung haben eine Vorrichtung und ein Verfahren zur Spracherkennung eine hohe Wahrscheinlichkeit, akustische Übereinstimmungen mit unbeabsichtigten Tönen oder gesprochenen Worten, die nicht für die Spracherkennungseinrichtung vorgesehen sind,

auszuschließen. Das heißt, durch die Annahme der Vertrauenswerte gemäß der Erfindung haben eine Vorrichtung und ein Verfahren zur Spracherkennung, die das akustische Modell mit der besten Übereinstimmung mit einem Ton kennzeichnen, eine hohe Wahrscheinlichkeit, das am besten übereinstimmende akustische Modell auszuschließen, falls der Ton unbeabsichtigt oder nicht für die Spracherkennungseinrichtung vorgesehen ist, und eine hohe Wahrscheinlichkeit, das am besten übereinstimmende akustische Modell anzunehmen, falls der Ton ein oder mehrere Worte darstellt, die für die Spracherkennungseinrichtung vorgesehen sind.

Kurze Beschreibung der Zeichnungen

Figur 1 ist ein Blockschaltbild eines Beispiels einer Spracherkennungsvorrichtung gemäß der Erfindung.

Figur 2 zeigt schematisch ein Beispiel eines akustischen Befehlsmodells.

Figur 3 zeigt schematisch ein Beispiel eines akustischen Schweigemodells.

Figur 4 zeigt schematisch ein Beispiel des akustischen Schweigemodells von Figur 3, das mit dem Ende des akustischen Befehlsmodells von Figur 2 verkettet ist.

Figur 5 zeigt schematisch die Status und möglichen Übergänge zwischen Status für das kombinierte akustische Modell von Figur 4 zu jedem aus einer Anzahl von Zeitpunkten t.

Figur 6 ist ein Blockschaltbild eines Beispiels des Akustikprozessors von Figur 1.

Beschreibung der bevorzugten Ausführungsformen

Mit Bezugnahme auf Figur 1 umfasst die Spracherkennungsvorrichtung gemäß der Erfindung einen Akustikprozessor 10 zum Messen des Wertes von mindestens einem Merkmal von jedem aus einer Folge von mindestens zwei Tönen. Der Akustikprozessor 10 misst den Wert des Merkmals jedes Tons während jedes aus einer Reihe aufeinanderfolgender Zeitintervalle, um eine Reihe von Merkmalsignalen zu erzeugen, die die Merkmalwerte des Tons darstellen.

Wie unten ausführlicher beschrieben wird, kann der Akustikprozessor beispielsweise die Amplitude jedes Tons in einem oder mehreren Frequenzbändern während einer Folge von Zeitintervallen von zehn Millisekunden messen, um eine Folge von Merkmalvektorsignalen zu erzeugen, die die Amplitudenwerte des Tons darstellen. Bei Bedarf können die Merkmalvektorsignale quantisiert werden, indem jedes Merkmalvektorsignal durch ein Prototypvektorsignal aus einem Satz von Prototypvektorsignalen ersetzt wird, das am besten mit dem Merkmalvektorsignal übereinstimmt. Jedes Prototypvektorsignal hat eine Kennzeichnung, und folglich erzeugt der Akustikprozessor in diesem Fall eine Reihe von Kennzeichnungssignalen, die die Merkmalwerte des Tons darstellen.

Die Spracherkennungsvorrichtung umfasst außerdem einen Speicher 12 für akustische Befehlsmodelle zur Speicherung eines Satzes akustischer Befehlsmodelle. Jedes akustische

Befehlsmodell stellt eine oder mehrere Folgen akustischer Merkmalwerte dar, die eine Äußerung eines dem akustischen Befehlsmodell zugeordneten Befehls darstellen.

Die gespeicherten akustischen Befehlsmodelle können beispielsweise Markow-Modelle oder andere dynamische Programmiermodelle sein. Die Parameter der akustischen Befehlsmodelle können aus einem bekannten Übungstext geschätzt werden, beispielsweise durch Glättungsparameter, die durch den Vorwärts-Rückwärts-Algorithmus erhalten werden. (Siehe beispielsweise F. Jelinek, "Continous Speech Recognition by Statistical Methods." Proceedings of the IEEE, Band 64, Nr. 4, April 1976, Seiten 532 bis 556.)

Vorzugsweise stellt jedes akustische Befehlsmodell einen isolierten, gesprochenen Befehl dar (das heißt, unabhängig vom Kontext früherer und nachfolgender Äußerungen). Kontextunabhängige akustische Befehlsmodelle können beispielsweise manuell aus Modellen von Phonemen oder automatisch erzeugt werden, beispielsweise durch das von Lalit R. Bahl et al. in der US-Patentschrift 4 759 068, mit dem Titel "Constructing Markov Models of Words From Multiple Utterances", beschriebene Verfahren oder durch jedes andere bekannte Verfahren zur Erzeugung kontextunabhängiger Modelle.

Alternativ können kontextabhängige Modelle aus kontextunabhängigen Modellen erzeugt werden, indem Äußerungen eines Befehls in kontextabhängige Kategorien gruppiert werden. Ein Kontext kann zum Beispiel manuell oder automatisch ausgewählt werden, indem jedes einem Befehl entsprechende Merkmalsignal mit seinem Kontext gekennzeichnet wird und indem die Merkmalsignale gemäß ihrem Kontext gruppiert werden, um

eine ausgewählte Bewertungsfunktion zu optimieren. (Siehe beispielsweise Lalit R. Bahl et al., "Apparatus and Method of Grouping Utterances of a Phoneme into Context-Dependent Categories Based on Sound-Similarity for Automatic Speech Recognition.", US-Patentschrift 5 195 167.)

Figur 2 zeigt schematisch ein Beispiel eines hypothetischen akustischen Befehlsmodells. In diesem Beispiel umfasst das akustische Befehlsmodell vier Status S1, S2, S3 und S4, die in Figur 2 als Punkte dargestellt werden. Das Modell beginnt beim Anfangsstatus S1 und endet beim letzten Status S4. Die gestrichelten Nullübergänge bedeuten, dass kein akustisches Merkmalsignal vom Akustikprozessor 10 ausgegeben wurde. Jedem Übergang mit durchgezogener Linie entspricht eine Ausgabewahrscheinlichkeitsverteilung über alle vom Akustikprozessor 10 erzeugten Merkmalvektorsignale oder Kennzeichnungssignale. Für jeden Status des Modells gibt es eine entsprechende Wahrscheinlichkeitsverteilung über die Übergänge aus diesem Status heraus.

Wiederum mit Bezugnahme auf Figur 1 umfasst die Spracherkennungsvorrichtung außerdem einen Vergleichswertprozessor 14 zum Erzeugen eines Vergleichswertes für jeden Ton und ein oder mehrere akustische Befehlsmodelle aus dem Satz akustischer Befehlsmodelle im Speicher 12 für akustische Befehlsmodelle. Jeder Vergleichswert umfasst eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Befehlsmodell und einer Folge dem Ton entsprechender Merkmalsignale vom Akustikprozessor 10.

Ein Erkennungsschwellenkomparator und -ausgabemittel 16 gibt ein Erkennungssignal aus, das dem Befehlsmodell aus dem

Speicher 12 für akustische Befehlsmodelle mit dem besten Vergleichswert für einen aktuellen Ton entspricht, falls der beste Vergleichswert für den aktuellen Ton besser als ein Erkennungsschwellenwert für den aktuellen Ton ist. Die Erkennungsschwelle für den aktuellen Ton umfasst einen ersten Vertrauenswert aus dem Speicher 18 für Vertrauenswerte, falls der beste Vergleichswert für einen früheren Ton besser als eine Erkennungsschwelle für diesen früheren Ton war. Die Erkennungsschwelle für den aktuellen Ton umfasst einen zweiten Vertrauenswert aus dem Speicher 18 für Vertrauenswerte, der besser als der erste Vertrauenswert ist, falls der beste Vergleichswert für einen früheren Ton schlechter als die Erkennungsschwelle für diesen früheren Ton war.

Die Spracherkennungsvorrichtung kann außerdem einen Speicher 20 für akustische Schweigemodelle zur Speicherung von mindestens einem akustischem Schweigemodell, das eine oder mehrere Folgen akustischer Merkmalwerte darstellt, die das Fehlen einer gesprochenen Äußerung darstellen. Das akustische Schweigemodell kann beispielsweise ein Markow-Modell oder ein anderes dynamisches Programmiermodell sein. Die Parameter des akustischen Schweigemodells können aus einem bekannten geäußerten Übungstext beispielsweise durch Glättungsparameter geschätzt werden, die auf dieselbe Weise wie bei den akustischen Befehlsmodellen aus dem Vorwärts-Rückwärts-Algorithmus erhalten werden.

Figur 3 zeigt schematisch ein Beispiel eines akustischen Schweigemodells. Das Modell beginnt beim Anfangsstatus S4 und endet beim Endstatus S10. Die gestrichelten Nullübergänge bedeuten, dass kein akustisches Merkmalsignal ausgegeben wird. Jedem Übergang mit durchgezogener Linie entspricht eine

Ausgabewahrscheinlichkeitsverteilung über die vom Akustikprozessor 10 erzeugten Merkmalsignale (zum Beispiel Merkmalvektorsignale oder Kennzeichnungssignale). Für jeden Status S4 bis S10 gibt es eine entsprechende Wahrscheinlichkeitsverteilung über die Übergänge aus diesem Status heraus.

Wiederum mit Bezugnahme auf Figur 1 erzeugt der Vergleichswertprozessor 14 einen Vergleichswert für jeden Ton und das akustische Schweigemodell im Speicher 20 für akustische Schweigemodelle. Jeder Vergleichswert mit dem akustischen Schweigemodell umfasst eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Schweigemodell und einer Folge dem Ton entsprechender Merkmalsignale.

In dieser Variante der Erfindung umfasst die vom Erkennungsschwellenkomparator und -ausgabemittel 16 verwendete Erkennungsschwelle den ersten Vertrauenswert, falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als eine aus dem Speicher 22 für Schweigevergleichs- und Schweigedauerschwellen erhaltene Schweigevergleichsschwelle ist, und falls der frühere Ton eine Dauer hat, die eine im Speicher 22 für Schweigevergleichs- und Schweigedauerschwellen gespeicherte Schweigedauerschwelle überschreitet. Alternativ umfasst die Erkennungsschwelle für den aktuellen Ton den ersten Vertrauenswert, falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die geringer als die Schweigedauerschwelle ist, und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches

Befehlsmodell besser als eine Erkennungsschwelle für diesen nächsten früheren Ton war. Schließlich umfasst die Erkennungsschwelle für den aktuellen Ton den ersten Vertrauenswert, falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell besser als eine Erkennungsschwelle für diesen früheren Ton war.

In dieser Ausführungsform der Erfindung umfasst die Erkennungsschwelle für den aktuellen Ton den zweiten Vertrauenswert, der besser als der erste Vertrauenswert aus dem Speicher 18 für Vertrauenswerte ist, falls der Vergleichswert vom Vergleichswertprozessor 18 für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die geringer als die Schweigedauerschwelle ist, und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen nächsten früheren Ton war. Alternativ umfasst die Erkennungsschwelle für den aktuellen Ton den zweiten Vertrauenswert, der besser als der erste Vertrauenswert ist, falls der Vergleichswert den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen früheren Ton war.

Zur Erzeugung eines Vergleichswertes für jeden Ton und jedes von einem oder mehreren akustischen Befehlsmodellen aus dem

Satz akustischer Befehlsmodelle im Speicher 12 für akustische Befehlsmodelle und zur Erzeugung eines Vergleichswertes für jeden Ton und das akustische Schweigemodell im Speicher 20 für akustische Schweigemodelle kann das akustische Schweigemodell von Figur 3 mit dem Ende des akustischen Befehlsmodells von Figur 2 verkettet werden, wie in Figur 4 gezeigt wird. Das kombinierte Modell beginnt im Anfangsstatus S1 und endet im Endstatus S10.

Die Status S1 bis S10 und die möglichen Übergänge zwischen den Status für das kombinierte akustische Modell von Figur 4 werden zu jedem aus einer Anzahl von Zeitpunkten t in Figur 5 schematisch gezeigt. Für jedes der Zeitintervalle zwischen t=n-1 und t=n erzeugt der Akustikprozessor ein Merkmalsignal X_n.

Für jeden Status des in Figur 4 gezeigten kombinierten Modells wird die bedingte Wahrscheinlichkeit $P(s_t = S_\sigma \mid X_1 \ldots X_t)$, dass der Status s_t zum Zeitpunkt t unter Berücksichtigung des Auftretens von Merkmalsignalen X_1 bis X_t, die zu den Zeitpunkten 1 bis t jeweils vom Akustikprozessor 10 erzeugt werden, gleich dem Status S_σ ist, durch die Gleichungen 1 bis 10 erhalten.

$$P(s_t = S1 \mid X_1 \ldots X_t) = \text{“}P(s_{t-1} = S1)\ P(s_t = S1 \mid s_{t-1} = S1)\ P(X_t \mid s_t = S1, s_{t-1} = S1)\text{”} \quad [1]$$

$$\begin{aligned} P(s_t = S2 \mid X_1 \ldots X_t) = {} & \text{“}P(s_{t-1} = S1)\ P(s_t = S2 \mid s_{t-1} = S1)\ P(X_t \mid s_t = S2, s_{t-1} = S1)\text{”} \\ & + P(s_t = S1)\ P(s_t = S2 \mid s_t = S1) \\ & + \text{“}P(s_{t-1} = S2)\ P(s_t = S2 \mid s_{t-1} = S2)\ P(X_t \mid s_t = S2, s_{t-1} = S2)\text{”} \end{aligned} \quad [2]$$

$$P(s_t = S3 \mid X_1 \ldots X_t) = \Pi P(s_{t-1} = S2)\ P(s_t = S3 \mid s_{t-1} = S2)\ P(X_t \mid s_t = S3, s_{t-1} = S2)''$$
$$+\ P(s_t = S2)\ P(s_t = S3 \mid s_t = S2)$$
$$+\ \Pi P(s_{t-1} = S3)\ P(s_t = S3 \mid s_{t-1} = S3)\ P(X_t \mid s_t = S3, s_{t-1} = S3)'' \quad [3]$$

$$P(s_t = S4 \mid X_1 \ldots X_t) = \Pi P(s_{t-1} = S3)\ P(s_t = S4 \mid s_{t-1} = S3)\ P(X_t \mid s_t = S4, s_{t-1} = S3)''$$
$$+\ P(s_t = S3)\ P(s_t = S4 \mid s_t = S3) \quad [4]$$

$$P(s_t = S6 \mid X_1 \ldots X_t) = \Pi P(s_{t-1} = S5)\ P(s_t = S6 \mid s_{t-1} = S5)\ P(X_t \mid s_t = S6, s_{t-1} = S5)''$$
$$+\ \Pi P(s_{t-1} = S6)\ P(s_t = S6 \mid s_{t-1} = S6)\ P(X_t \mid s_t = S6, s_{t-1} = S6)'' \quad [5]$$

$$P(s_t = S6 \mid X_1 \ldots X_t) = \Pi P(s_{t-1} = S5)\ P(s_t = S6 \mid s_{t-1} = S5)\ P(X_t \mid s_t = S6, s_{t-1} = S5)''$$
$$+\ \Pi P(s_{t-1} = S6)\ P(s_t = S6 \mid s_{t-1} = S6)\ P(X_t \mid s_t = S6, s_{t-1} = S6)'' \quad [6]$$

$$P(s_t = S7 \mid X_1 \ldots X_t) = \Pi P(s_{t-1} = S6)\ P(s_t = S7 \mid s_{t-1} = S6)\ P(X_t \mid s_t = S7, s_{t-1} = S6)''$$
$$+\ P(s_{t-1} = S7)\ P(s_t = S7 \mid s_{t-1} = S7)\ P(X_t \mid s_t = S7, s_{t-1} = S7)'' \quad [7]$$

$$P(s_t = S8 \mid X_1 \ldots X_t) = \Pi P(s_{t-1} = S4)\ P(s_t = S8 \mid s_{t-1} = S4)\ P(X_t \mid s_t = S8, s_{t-1} = S4)'' \quad [8]$$

$$P(s_t = S9 \mid X_1 \ldots X_t) = \text{"}\Sigma P(s_{t-1} = S8)\ P(s_t = S9 \mid s_{t-1} = S8)\ P(X_t \mid s_t = S9, s_{t-1} = S8)\text{"} \quad [9]$$

$$\begin{aligned} P(s_t = S10 \mid X_1 \ldots X_t) = {} & P(s_t = S4)\ P(s_t = S10 \mid s_t = S4) \\ & + P(s_t = S8)\ P(s_t = S10 \mid s_t = S8) \\ & + P(s_t = S9)\ P(s_t = S10 \mid s_t = S9) \\ & + \text{"}\Sigma P(s_{t-1} = S7)\ P(s_t = S10 \mid s_{t-1} = S7) \\ & \quad P(X_t \mid s_t = S10, s_{t-1} = S7)\text{"} \\ & + \text{"}\Sigma P(s_{t-1} = S9)\ P(s_t = S10 \mid s_{t-1} = S9) \\ & \quad P(X_t \mid s_t = S10, s_{t-1} = S9)\text{"} \end{aligned} \quad [10]$$

Zur Normierung der bedingten Statuswahrscheinlichkeiten, um die verschiedenen Anzahlen von Merkmalsignalen ($X_1 \ldots X_n$) zu verschiedenen Zeitpunkten t zu berücksichtigen, kann ein normierter Statusausgabewert Q für einen Status σ zum Zeitpunkt t durch die Gleichung 11 gegeben werden.

$$Q(\sigma, t) = \frac{P(s_t = S\sigma \mid X_1 \ldots X_t)}{\prod_{i=1}^{t} P(X_i)} \quad [11]$$

Geschätzte Werte für die bedingten Wahrscheinlichkeiten $P(s_t = S_\sigma \mid X_1 \ldots X_t)$ der Status (in diesem Beispiel der Status S1 bis S10) können aus den Gleichungen 1 bis 10 erhalten werden, indem die Werte der Übergangswahrscheinlichkeitsparameter und der Ausgabewahrscheinlichkeitsparameter der akustischen Befehlsmodelle und der akustischen Schweigemodelle verwendet werden.

Geschätzte Werte für den normierten Statusausgabewert Q können aus der Gleichung 11 erhalten werden, indem die Wahrscheinlichkeit $P(X_i)$ jedes beobachteten Merkmalsignals X_i als Produkt aus der bedingten Wahrscheinlichkeit $P(X_i \mid X_{i-1})$ des Merkmalsignals X_i unter Berücksichtigung des unmittelbar früheren Auftretens des Merkmalsignals X_{i-1}, multipliziert mit der Wahrscheinlichkeit $P(X_{i-1})$ des Auftretens des Merkmalsignals X_{i-1}, geschätzt wird. Der Wert von $P(X_i \mid X_{i-1})$ $P(X_{i-1})$ kann für alle Merkmalsignale X_i und X_{i-1} geschätzt werden, indem das Auftreten von Merkmalsignalen gezählt wird, die gemäß der Gleichung 12 aus einem Übungstext erzeugt werden.

$$P(X_i|X_{i-1})P(X_{i-1}) = \frac{N(X_i, X_{i-1})}{N(X_{i-1})} \frac{N(X_{i-1})}{N} = \frac{N(X_i, X_{i-1})}{N} \qquad [12]$$

In der Gleichung 12 ist $N(X_i , X_{i-1})$ die Anzahl des Auftretens des Merkmalsignals X_i , dem das durch die Äußerung des Trainingstextes erzeugte Merkmalsignal X_{i-1} unmittelbar vorangeht, und N ist die Gesamtanzahl von Merkmalsignalen, die durch die Äußerung des Übungstextes erzeugt werden.

Aus der obigen Gleichung 11 können die normierten Statusausgabewerte Q(S4, t) und Q(S10, t) für die Status S4 und S10 des kombinierten Modells von Figur 4 erhalten werden. Der Status S4 ist der letzte Status des Befehlsmodells und der erste Status des Schweigemodells. Der Status S10 ist der letzte Status des Schweigemodells.

In einem Beispiel der Erfindung kann ein Vergleichswert für einen Ton und das akustische Schweigemodell zum Zeitpunkt t durch das Verhältnis des normierten Statusausgabewertes Q[S10,t] für den Status S10 dividiert durch den normierten Statusausgabewert Q[S4,t] für den Status S4 gegeben werden, wie in der Gleichung 13 gezeigt wird.

$$\text{Schweigestart-Vergleichswert} = \frac{Q[S10,t]}{Q[S4,t]} \qquad [13]$$

Der Zeitpunkt $t = t_{start}$, zu dem der Vergleichswert für den Ton und das akustische Schweigemodell (Gleichung 13) zuerst eine Schweigevergleichsschwelle überschreitet, kann als der Beginn eines Schweigeintervalls betrachtet werden. Die Schweigevergleichsschwelle ist ein Abgleichparameter, der vom Benutzer eingestellt werden kann. Es wurde festgestellt, dass eine Schweigevergleichsschwelle von 10^{15} gute Ergebnisse erzeugt.

Das Ende des Schweigeintervalls kann beispielsweise festgestellt werden, indem das Verhältnis des normierten Statusausgabewertes Q[S10,t] für den Status S10 zum Zeitpunkt t, dividiert durch den erhaltenen Maximalwert für den normierten Statusausgabewert $Q_{max}[S10, t_{start}, \ldots t]$ für den Status S10 über die Zeitintervalle t_{start} bis t ausgewertet wird.

$$\text{Schweigeende-Vergleichswert} = \frac{Q[S10,t]}{Q_{max}[S10, t_{start}, \ldots t]} \qquad [14]$$

Der Zeitpunkt t = t_{End}, zu dem der Wert des Schweigeende-Vergleichswertes von Gleichung 14 zuerst unter den Wert einer Schweigeendeschwelle fällt, kann als das Ende des Schweigeintervalls betrachtet werden. Der Wert der Schweigeendeschwelle ist ein Abgleichparameter, der vom Benutzer eingestellt werden kann. Es wurde festgestellt, dass ein Wert von 10^{-25} gute Ergebnisse bereitstellt.

Falls der Vergleichswert für den Ton und das akustische Schweigemodell, wie er durch die Gleichung 13 gegeben wird, besser als die Schweigevergleichsschwelle ist, wird das Schweigen als beim ersten Zeitpunkt t_{start} beginnend betrachtet, zu dem das Verhältnis von Gleichung 13 die Schweigevergleichsschwelle überschreitet. Das Schweigen wird als beim Zeitpunkt t_{End} endend betrachtet, zu dem das Verhältnis von Gleichung 14 kleiner als der zugeordnete Abgleichparameter ist. Die Dauer des Schweigens ist dann (t_{End} - t_{start}).

Für die Entscheidung, ob die Erkennungsschwelle der erste Vertrauenswert oder der zweite Vertrauenswert sein sollte, ist die im Speicher 22 für Schweigevergleichs- und Schweigedauerschwellen gespeicherte Schweigedauerschwelle ein Abgleichparameter, der vom Benutzer eingestellt werden kann. Es wurde beispielsweise festgestellt, dass eine Schweigedauerschwelle von 25 Zentisekunden gute Ergebnisse bereitstellt.

Der Vergleichswert für jeden Ton und ein akustisches Befehlsmodell, das den Status S1 bis S4 der Figuren 2 und 4 entspricht, kann folgendermaßen erhalten werden. Falls das Verhältnis von Gleichung 13 die Schweigevergleichsschwelle

nicht vor dem Zeitpunkt t_{End} überschreitet, kann der Vergleichswert für jeden Ton und das den Status S1 bis S4 der Figuren 2 und 4 entsprechende akustische Befehlsmodell durch den maximalen normierten Statusausgabewert Q[S10, t'_{End}, ...t_{End}] für den Status S10 über die Zeitintervalle t'_{End} bis t_{End} gegeben werden, wobei t'_{End} das Ende des vorhergehenden Tons oder Schweigens ist und wobei t_{End} das Ende des aktuellen Tons oder Schweigens ist. Alternativ kann der Vergleichswert für jeden Ton und das akustische Befehlsmodell durch die Summe der normierten Statusausgabewerte Q[S10, t] für den Status S10 über die Zeitintervalle t'_{End} bis t_{End} gegeben werden.

Falls jedoch das Verhältnis von Gleichung 13 die Schweigevergleichsschwelle vor dem Zeitpunkt t_{End} überschreitet, kann der Vergleichswert für den Ton und das akustische Befehlsmodell durch den normierten Statusausgabewert Q[S4, t_{Start}] für den Status S4 zum Zeitpunkt t_{Start} gegeben werden. Alternativ kann der Vergleichswert für jeden Ton und das akustische Befehlsmodell durch die Summe aus den normierten Statusausgabewerten Q[S4, t] für den Status S4 über die Zeitintervalle t'_{End} bis t_{Start} gegeben werden.

Der erste Vertrauenswert und der zweite Vertrauenswert für die Erkennungsschwelle sind Abgleichparameter, die vom Benutzer eingestellt werden können. Die ersten und zweiten Vertrauenswerte können beispielsweise folgendermaßen erzeugt werden.

Ein Übungstext, der im Vokabular enthaltene Befehlsworte, die durch gespeicherte akustische Befehlsmodelle dargestellt werden, und außerdem nicht im Vokabular enthaltene Worte umfasst, die nicht durch gespeicherte akustische

Befehlsmodelle dargestellt werden, wird von einem oder mehreren Sprechern gesprochen. Unter Verwendung der Spracherkennungsvorrichtung gemäß der Erfindung, jedoch ohne eine Erkennungsschwelle, wird eine Folge erkannter Worte erzeugt, die am besten mit dem gesprochenen, bekannten Übungstext übereinstimmen. Jedem von der Spracherkennungsvorrichtung ausgegebenen Wort oder Befehl wird ein Vergleichswert zugeordnet.

Durch den Vergleich der Befehlsworte im bekannten Übungstext mit den von der Spracherkennungsvorrichtung ausgegebenen, erkannten Worten können korrekt erkannte Worte und falsch erkannte Worte gekennzeichnet werden. Der erste Vertrauenswert kann beispielsweise der beste Vergleichswert sein, der schlechter als die Vergleichswerte von 99 % bis 100 % der korrekt erkannten Worte ist. Der zweite Vertrauenswert kann beispielsweise der schlechteste Vergleichswert sein, der besser als die Vergleichswerte von beispielsweise 99 % bis 100 % der falsch erkannten Worte im Übungstext ist.

Das vom Erkennungsschwellenkomparator und -ausgabemittel 16 ausgegebene Erkennungssignal kann ein Befehlssignal zum Aufrufen eines dem Befehl zugeordneten Programms umfassen. Das Befehlssignal kann beispielsweise die manuelle Eingabe von einem Befehl entsprechenden Tastenanschlägen simulieren. Alternativ kann das Befehlssignal ein Anwendungsprogramm-Schnittstellenaufruf sein.

Das Erkennungsschwellenkomparator und -ausgabemittel 16 kann eine Anzeige, beispielsweise eine Kathodenstrahlröhre, eine Flüssigkristallanzeige oder einen Drucker umfassen. Das Erkennungsschwellenkomparator und -ausgabemittel 16 kann ein

oder mehrere Worte anzeigen, die dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entsprechen, falls der beste Vergleichswert für den aktuellen Ton besser als der Erkennungsschwellenwert für den aktuellen Ton ist.

Das Ausgabemittel 16 kann wahlweise ein Signal für einen nicht erkennbaren Ton ausgeben, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist. Die Ausgabe 16 kann beispielsweise eine Anzeige für einen nicht erkennbaren Ton anzeigen, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist. Die Anzeige für einen nicht erkennbaren Ton kann ein oder mehrere angezeigte Fragezeichen umfassen.

Jeder vom Akustikprozessor 10 gemessene Ton kann ein Vokalton oder ein anderer Ton sein. Jeder einem akustischen Befehlsmodell zugeordnete Befehl umfasst vorzugsweise mindestens ein Wort.

Zu Beginn einer Spracherkennungssitzung kann die Erkennungsschwelle am ersten Vertrauenswert oder am zweiten Vertrauenswert initialisiert werden. Vorzugsweise wird die Erkennungsschwelle für den aktuellen Ton zu Beginn einer Spracherkennungssitzung am ersten Vertrauenswert initialisiert.

Die Spracherkennungsvorrichtung gemäß der vorliegenden Erfindung kann mit jeder bestehenden Spracherkennungseinrichtung verwendet werden, beispielsweise mit dem IBM Speech Server Series- (Warenzeichen) Produkt. Der Vergleichswertprozessor 14 und das

Erkennungsschwellenkomparator und -ausgabemittel 16 können beispielsweise geeignet programmierte spezielle oder allgemeine digitale Prozessoren sein. Der Speicher 12 für akustische Befehlsmodelle, der Speicher 18 für Vertrauenswerte, der Speicher 20 für akustische Schweigemodelle und der Speicher 22 für Schweigevergleichs- und Schweigedauerschwellen können beispielsweise einen elektronisch lesbaren Computerspeicher umfassen.

Ein Beispiel des Akustikprozessors 10 von Figur 3 wird in Figur 6 gezeigt. Der Akustikprozessor umfasst ein Mikrofon 24 zum Erzeugen eines der Äußerung entsprechenden, analogen elektrischen Signals. Das analoge elektrische Signal vom Mikrofon 24 wird durch den Analog-Digital-Umsetzer 26 in ein digitales elektrisches Signal umgesetzt. Zu diesem Zweck kann das analoge Signal beispielsweise bei einer Geschwindigkeit von zwanzig Kilohertz vom Analog-Digital-Umsetzer 26 abgetastet werden.

Ein Fenstergenerator 28 erhält beispielsweise alle zehn Millisekunden (eine Zentisekunde) einen Abtastwert des digitalen Signals mit einer Dauer von zwanzig Millisekunden vom Analog-Digital-Umsetzer 26. Jeder zwanzig Millisekunden lange Abtastwert des digitalen Signals wird vom Spektrumanalysator 30 analysiert, um die Amplitude des digitalen Signalabtastwertes in jedem der beispielsweise zwanzig Frequenzbänder zu erhalten. Vorzugsweise erzeugt der Spektrumanalysator 30 außerdem ein einundzwanzigdimensionales Signal, das die Gesamtamplitude oder Gesamtleistung des zwanzig Millisekunden langen digitalen Signalabtastwertes darstellt. Der Spektrumanalysator 30 kann beispielsweise ein

schneller Fourier-Transformations-Prozessor sein. Alternativ kann er eine Gruppe von zwanzig Bandpassfiltern sein.

Die vom Spektrumanalysator 30 erzeugten einundzwanzigdimensionalen Vektorsignale können so bearbeitet werden, dass Hintergrundrauschen durch einen adaptiven Rauschunterdrückungsprozessor 32 entfernt wird. Der Rauschunterdrückungsprozessor 32 subtrahiert einen Rauschvektor N(t) von dem in den Rauschunterdrückungsprozessor eingegebenen Merkmalvektor F(t), um einen ausgegebenen Merkmalvektor F'(t) zu erzeugen. Der Rauschunterdrückungsprozessor 32 passt sich an ändernde Rauschpegel an, indem er den Rauschvektor N(t) jedesmal, wenn der frühere Merkmalvektor F(t -1) als Rauschen oder Schweigen gekennzeichnet wird, periodisch aktualisiert. Der Rauschvektor N(t) wird gemäß der folgenden Formel aktualisiert

$$N(t) = \frac{N(t-1) + k[F(t-1) - Fp(t-1)]}{(1+k)}, \qquad [15]$$

wobei N(t) der Rauschvektor zum Zeitpunkt t, N(t - 1) der Rauschvektor zum Zeitpunkt (t -1), k ein feststehender Parameter des adaptiven Rauschunterdrückungsmodells, F(t -1) der in den Rauschunterdrückungsprozessor 32 eingegebene Merkmalvektor zum Zeitpunkt (t - 1) ist und der Rauschen oder Schweigen darstellt, und Fp(t -1) ein Schweige- oder Rauschprototypvektor aus dem Speicher 24 ist, der die größte Annäherung zum Merkmalvektor F(t -1) hat.

Der frühere Merkmalvektor F(t - 1) wird als Rauschen oder Schweigen erkannt, falls (a) die Gesamtenergie des Vektors unter einer Schwelle liegt oder (b) der Prototypvektor im Anpassungsprototypvektorspeicher 36 mit der größten Annäherung an den Merkmalvektor ein Prototyp ist, der Rauschen oder Schweigen darstellt. Für die Analyse der Gesamtenergie des Merkmalvektors kann die Schwelle beispielsweise das fünfte Percentil aller Merkmalvektoren sein (sowohl Sprache als auch Schweigen entsprechend), die in den beiden Sekunden vor der Auswertung des Merkmalvektors erzeugt werden.

Nach der Rauschunterdrückung wird der Merkmalvektor F'(t) zur Anpassung an Änderungen der Lautstärke der eingegebenen Sprache durch den Normierungsprozessor 38 für kurzzeitige Mittelwerte normiert. Der Normierungsprozessor 38 normiert den einundzwanzigdimensionalen Merkmalvektor F'(t), um einen normierten einundzwanzigdimensionalen Merkmalvektor X(t) zu erzeugen. Die einundzwanzigste Dimension des Merkmalvektors F'(t), die die Gesamtamplitude oder die Gesamtenergie darstellt, wird gelöscht. Jede Komponente i des normierten Merkmalvektors X(t) zum Zeitpunkt t kann beispielsweise durch die folgende Gleichung im logarithmischen Bereich gegeben werden

$$X_i(t) = F_i(t) - Z(t) \qquad [16]$$

wobei $F'_i(t)$ die i-te Komponente des nicht normierten Vektors zum Zeitpunkt t ist und wobei Z(t) ein gewichtetes Mittel der Komponenten von F'(t) und Z(t -1) gemäß den Gleichungen 17 und 18 ist:

$$Z(t) = 0.9Z(t-1) + 0.1M(t) \qquad [17]$$

und wobei

$$M(t) = \frac{1}{20}\sum_i F_i(t) \qquad [18]$$

Der normierte einundzwanzigdimensionale Merkmalvektor X(t) kann außerdem zur Anpassung an Änderungen bei der Aussprache von Sprachtönen durch eine adaptive Kennzeichnungseinrichtung 40 verarbeitet werden. Ein angepasster einundzwanzigdimensionaler Merkmalvektor X'(t) wird erzeugt, indem ein einundzwanzigdimensionaler Anpassungsvektor A(t) vom einundzwanzigdimensionalen Merkmalvektor X(t), der zum Eingang der adaptiven Kennzeichnungseinrichtung 40 gesendet wird, subtrahiert wird. Der Anpassungsvektor A(t) zum Zeitpunkt t kann beispielsweise durch die folgende Formel gegeben werden

$$A(t) = \frac{A(t-1) + k[X(t-1) - Xp(t-1)]}{(1+k)} . \qquad [19]$$

wobei k ein feststehender Parameter des adaptiven Kennzeichnungsmodells, X(t - 1) der zum Zeitpunkt (t -1) in die adaptive Kennzeichnungseinrichtung 40 eingegebene, normierte einundzwanzigdimensionale Vektor, Xp(t - 1) der Anpassungsprototypvektor (aus dem Anpassungsprototypspeicher

36) mit der größten Annäherung an den einundzwanzigdimensionalen Merkmalvektor X(t - 1) zum Zeitpunkt (t - 1) und A(t - 1) der Anpassungsvektor zum Zeitpunkt (t - 1) ist.

Das angepasste einundzwanzigdimensionale Merkmalvektorsignal X'(t) aus der adaptiven Kennzeichnungseinrichtung 40 wird vorzugsweise zu einem Hörmodell (auditory model) 42 gesendet. Das Hörmodell 42 kann beispielsweise ein Modell davon bereitstellen, wie das menschliche Hörsystem Tonsignale wahrnimmt. Ein Beispiel eines Hörsystems wird in der US-Patentschrift 4 980 918 von Bahl et al. mit dem Titel "Speech Recognition System with Efficient Storage and Rapid Assembly of Phonological Graphs" beschrieben.

Vorzugsweise berechnet das Hörmodell 42 gemäß der Erfindung für jedes Frequenzband i des angepassten Merkmalvektorsignals X'(t) zum Zeitpunkt t einen neuen Parameter $E_i(t)$ gemäß den Gleichungen 20 und 21:

$$E_i(t) = K_1 + K_2(X'_i(t))(N_i(t-1)) \qquad [20]$$

wobei

$$N_i(t) = K_3 \times N_i(t-1) - E_i(t-1) \qquad [21]$$

und wobei K_1, K_2 und K_3 feststehende Parameter des Hörmodells sind.

Für jedes Zentisekunden-Zeitintervall ist die Ausgabe des Hörmodells 42 ein geändertes einundzwanzigdimensionales Merkmalvektorsignal. Dieser Merkmalvektor wird durch eine einundzwanzigste Dimension mit einem Wert, der gleich der Quadratwurzel aus der Summe der Quadrate der anderen zwanzig Dimensionen ist, erhöht.

Für jedes Zentisekunden-Zeitintervall verkettet eine Verkettungseinrichtung 44 vorzugsweise neun einundzwanzigdimensionalen Merkmalvektoren, die das eine aktuelle Zentisekunden-Zeitintervall, die vier vorhergehenden Zentisekunden-Zeitintervalle und die vier folgenden Zentisekunden-Zeitintervalle darstellen, um einen einzigen verknüpften Vektor von 189 Dimensionen zu bilden. Jeder verknüpfte Vektor der 189 Dimensionen wird vorzugsweise in einem Drehoperator 46 mit einer Drehmatrix multipliziert, um den verknüpften Vektor zu drehen und um den verknüpften Vektor auf fünfzig Dimensionen zu reduzieren.

Die im Drehoperator 46 verwendete Drehmatrix kann beispielsweise erhalten werden, indem ein Satz verknüpfter Vektoren von 189 Dimensionen, die während einer Trainingssitzung erhalten werden, in M Klassen eingeteilt werden. Die Kovarianzmatrix wird für alle der verknüpften Vektoren im Trainingssatz mit dem Inversen der in der Klasse enthaltenen Kovarianzmatrix für alle der verknüpften Vektoren in allen M Klassen multipliziert. Die ersten fünfzig Eigenvektoren der resultierenden Matrix bilden die Drehmatrix. (Siehe zum Beispiel "Vector Quantization Procedure For Speech

Recognition Systems Using Discrete Parameter Phoneme-Based Markov Word Models" von L.R. Bahl et al., IBM Technical Disclosure Bulletin, Band 32, Nr. 7, Dezember 1989, Seiten 320 und 321.)

Der Fenstergenerator 28, der Spektrumanalysator 30, der adaptive Rauschunterdrückungsprozessor 32, der Normierungsprozessor 38 für kurzzeitige Mittelwerte, die adaptive Kennzeichnungseinrichtung 40, das Hörmodell 42, die Verkettungseinrichtung 44 und der Drehoperator 46 können geeignet programmierte spezielle oder allgemeine digitale Signalprozessoren sein. Die Prototypspeicher 34 und 36 können elektronische Computerspeicher der oben erläuterten Typen sein.

Die Prototypvektoren im Prototypspeicher 34 können beispielsweise erhalten werden, indem die Merkmalvektorsignale aus einem Trainingssatz in eine Vielzahl von Zuordnungseinheiten eingeordnet und anschließend die Durchschnitts- und Standardabweichung für jede Zuordnungseinheit berechnet wird, um die Parameterwerte des Prototypvektors zu bilden. Wenn der Übungstext eine Folge von Wortsegmentmodellen (die ein Modell einer Folge von Worten bilden) und jedes Wortsegmentmodell eine Folge von Elementarmodellen mit angegebenen Positionen in den Wortsegmentmodellen umfasst, können die Merkmalvektorsignale in Gruppen geordnet werden, indem angegeben wird, dass jede Zuordnungseinheit einem einzigen Elementarmodell in einer einzigen Position in einem einzigen Wortsegmentmodell entspricht. Ein solches Verfahren wird in der US-Patentanmeldung mit der Seriennr. 730 714, eingereicht am 16. Juli 1991, mit dem Titel "Fast Algorithm for Deriving Acoustic

Prototypes for Automatic Speech Recognition" ausführlicher beschrieben.

Alternativ können alle akustischen Merkmalvektoren, die durch die Äußerung eines Übungstextes erzeugt werden und die einem gegebenen Elementarmodell entsprechen, durch euklidische K-Mittelwert-Zuordnung oder Gaußsche K-Mittelwert-Zuordnung oder beides in Gruppen eingeordnet werden. Ein solches Verfahren wird beispielsweise von Bahl et al. in der US-Patentschrift 5 182 773 mit dem Titel "Speaker-Independent Label Coding Apparatus" beschrieben.

A N S P R Ü C H E

1. Spracherkennungseinrichtung, die Folgendes umfasst:

einen Akustikprozessor (10) zum Messen des Wertes von mindestens einem Merkmal von jedem aus einer Folge von mindestens zwei Tönen, wobei der Akustikprozessor (10) den Wert des Merkmals jedes Tons während jedes aus einer Reihe aufeinanderfolgender Zeitintervalle misst, um eine Reihe von Merkmalsignalen zu erzeugen, die die Merkmalwerte des Tons darstellen;

Mittel (12) zum Speichern eines Satzes akustischer Befehlsmodelle, wobei jedes akustische Befehlsmodell eine oder mehrere Reihen akustischer Merkmalswerte darstellt, die eine Äußerung eines dem akustischen Befehlsmodell zugeordneten Befehls darstellen;

einen Vergleichswertprozessor (14) zum Erzeugen eines Vergleichswertes für jeden Ton und jedes von einem oder mehreren akustischen Befehlsmodellen aus dem Satz akustischer Befehlsmodelle, wobei jeder Vergleichswert eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Befehlsmodell und einer Reihe dem Ton entsprechender Merkmalsignale umfasst;

gekennzeichnet durch:

Mittel (16) zum Ausgeben eines Erkennungssignals, das dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entspricht, falls der beste Vergleichswert

für den aktuellen Ton besser als ein Erkennungsschwellenwert für den aktuellen Ton ist, wobei die Erkennungsschwelle für den aktuellen Ton Folgendes umfasst: (a) einen ersten Vertrauenswert, falls der beste Vergleichswert für einen früheren Ton besser als eine Erkennungsschwelle für diesen früheren Ton war, oder (b) einen zweiten Vertrauenswert, der besser als der erste Vertrauenswert ist, falls der beste Vergleichswert für einen früheren Ton schlechter als die Erkennungsschwelle für diesen früheren Ton war.

2. Spracherkennungsvorrichtung nach Anspruch 1, dadurch gekennzeichnet, dass der frühere Ton unmittelbar vor dem aktuellen Ton auftritt.

3. Spracherkennungsvorrichtung nach Anspruch 2, dadurch gekennzeichnet, dass:

die Vorrichtung außerdem Mittel (20) zum Speichern von mindestens einem akustischen Schweigemodell umfasst, das eine oder mehrere Reihen akustischer Merkmalswerte darstellt, die das Nichtvorhandensein einer gesprochenen Äußerung darstellen;

der Vergleichswertprozessor (10) für jeden Ton und das akustische Schweigemodell einen Vergleichswert erzeugt, wobei jeder Vergleichswert eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Schweigemodell und einer Reihe von dem Ton entsprechenden Merkmalsignalen umfasst; und

die Erkennungsschwelle für den aktuellen Ton den ersten Vertrauenswert umfasst, (a1) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als eine Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer aufweist, die eine Schweigedauerschwelle übersteigt, oder (a2) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die kürzer als die Schweigedauerschwelle ist und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell besser als eine Erkennungsschwelle für diesen nächsten früheren Ton war, oder (a3) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell besser als eine Erkennungsschwelle für diesen früheren Ton war; oder

dass die Erkennungsschwelle für den aktuellen Ton den zweiten Vertrauenswert umfasst, der besser als der erste Vertrauenswert ist, (b1) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die kürzer als die Schweigedauerschwelle ist, und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen nächsten früheren Ton war, oder (b2) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die

Schweigevergleichsschwelle ist und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen früheren Ton war.

4. Spracherkennungsvorrichtung nach Anspruch 3, dadurch gekennzeichnet, dass das Erkennungssignal ein Befehlssignal zum Aufrufen eines dem Befehl zugeordneten Programms umfasst.

5. Spracherkennungsvorrichtung nach Anspruch 4, dadurch gekennzeichnet, dass:

das Ausgabemittel (16) eine Anzeige umfasst; und

das Ausgabemittel (16) eines oder mehrere Worte anzeigt, die dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entsprechen, falls der beste Vergleichswert für den aktuellen Ton besser als der Erkennungsschwellenwert für den aktuellen Ton ist.

6. Spracherkennungsvorrichtung nach Anspruch 5, dadurch gekennzeichnet, dass das Ausgabemittel (16) ein Anzeigesignal für einen nicht erkennbaren Ton ausgibt, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist.

7. Spracherkennungsvorrichtung nach Anspruch 6, dadurch gekennzeichnet, dass das Ausgabemittel (16) eine Anzeige für einen nicht erkennbaren Ton anzeigt, falls der beste

Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist.

8. Spracherkennungsvorrichtung nach Anspruch 7, dadurch gekennzeichnet, dass die Anzeige für einen nicht erkennbaren Ton ein oder mehrere Fragezeichen umfasst.

9. Spracherkennungsvorrichtung nach Anspruch 1, dadurch gekennzeichnet, dass der Akustikprozessor (10) ein Mikrofon (24) umfasst.

10. Spracherkennungsvorrichtung nach Anspruch 1, dadurch gekennzeichnet, dass:

jeder Ton einen Vokalton umfasst; und

jeder Befehl mindestens ein Wort umfasst.

11. Spracherkennungsverfahren, das die folgenden Schritte umfasst:

Messen des Wertes von mindestens einem Merkmal von jedem aus einer Folge von mindestens zwei Tönen, wobei der Wert des Merkmals jedes Tons während jeder aus einer Reihe aufeinanderfolgender Zeitintervalle gemessen wird, um eine Reihe von Merkmalsignalen zu erzeugen, die die Merkmalwerte des Tons darstellen;

Speichern eines Satzes akustischer Befehlsmodelle, wobei jedes akustische Befehlsmodell eine oder mehrere Reihen akustischer Merkmalswerte darstellt, die eine Äußerung

eines dem akustischen Befehlsmodell zugeordneten Befehls darstellen;

Erzeugen eines Vergleichswertes für jeden Ton und jedes von einem oder mehreren akustischen Befehlsmodellen aus dem Satz akustischer Befehlsmodelle, wobei jeder Vergleichswert eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Befehlsmodell und einer Reihe dem Ton entsprechender Merkmalsignale umfasst;

gekennzeichnet durch

das Ausgeben eines Erkennungssignals, das dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entspricht, falls der beste Vergleichswert für den aktuellen Ton besser als ein Erkennungsschwellenwert für den aktuellen Ton ist, wobei die Erkennungsschwelle für den aktuellen Ton Folgendes umfasst: (a) ein erster Vertrauenswert, falls der beste Vergleichswert für einen früheren Ton besser als eine Erkennungsschwelle für diesen früheren Ton war, oder (b) ein zweiter Vertrauenswert, der besser als der erste Vertrauenswert ist, falls der beste Vergleichswert für einen früheren Ton schlechter als die Erkennungsschwelle für diesen früheren Ton war.

12. Spracherkennungsverfahren nach Anspruch 11, dadurch gekennzeichnet, dass der frühere Ton unmittelbar vor dem aktuellen Ton auftritt.

13. Spracherkennungsverfahren nach Anspruch 12, das außerdem die folgenden Schritte umfasst:

Speichern von mindestens einem akustischen Schweigemodell, das eine oder mehrere Reihen akustischer Merkmalswerte darstellt, die das Nichtvorhandensein einer gesprochenen Äußerung darstellen;

Erzeugen eines Vergleichswertes für jeden Ton und das akustische Schweigemodell, wobei jeder Vergleichswert eine Schätzung der Genauigkeit einer Übereinstimmung zwischen dem akustischen Schweigemodell und einer Reihe von dem Ton entsprechenden Merkmalsignalen umfasst; und das dadurch gekennzeichnet ist, dass

die Erkennungsschwelle für den aktuellen Ton den ersten Vertrauenswert umfasst, (a1) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als eine Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer aufweist, die eine Schweigedauerschwelle übersteigt, oder (a2) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die kürzer als die Schweigedauerschwelle ist und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell besser als eine Erkennungsschwelle für diesen nächsten früheren Ton war, oder (a3) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist und falls der beste Vergleichswert für den früheren Ton und ein akustisches

Befehlsmodell besser als eine Erkennungsschwelle für diesen früheren Ton war; oder dass die Erkennungsschwelle für den aktuellen Ton den zweiten Vertrauenswert umfasst, der besser als der erste Vertrauenswert ist, (b1) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell besser als die Schweigevergleichsschwelle ist und falls der frühere Ton eine Dauer hat, die kürzer als die Schweigedauerschwelle ist, und falls der beste Vergleichswert für den nächsten früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen nächsten früheren Ton war, oder (b2) falls der Vergleichswert für den früheren Ton und das akustische Schweigemodell schlechter als die Schweigevergleichsschwelle ist und falls der beste Vergleichswert für den früheren Ton und ein akustisches Befehlsmodell schlechter als die Erkennungsschwelle für diesen früheren Ton war.

14. Spracherkennungsverfahren nach Anspruch 13, dadurch gekennzeichnet, dass das Erkennungssignal ein Befehlssignal zum Aufrufen eines dem Befehl zugeordneten Programms umfasst.

15. Spracherkennungsverfahren nach Anspruch 14, das außerdem den Schritt des Anzeigens eines oder mehrerer Worte umfasst, die dem Befehlsmodell mit dem besten Vergleichswert für einen aktuellen Ton entsprechen, falls der beste Vergleichswert für den aktuellen Ton besser als der Erkennungsschwellenwert für den aktuellen Ton ist.

16. Spracherkennungsverfahren nach Anspruch 15, das außerdem den Schritt des Ausgebens eines Anzeigesignals für einen

nicht erkennbaren Ton umfasst, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist.

17. Spracherkennungsverfahren nach Anspruch 16, das außerdem den Schritt des Anzeigens einer Anzeige für einen nicht erkennbaren Ton umfasst, falls der beste Vergleichswert für den aktuellen Ton schlechter als der Erkennungsschwellenwert für den aktuellen Ton ist.

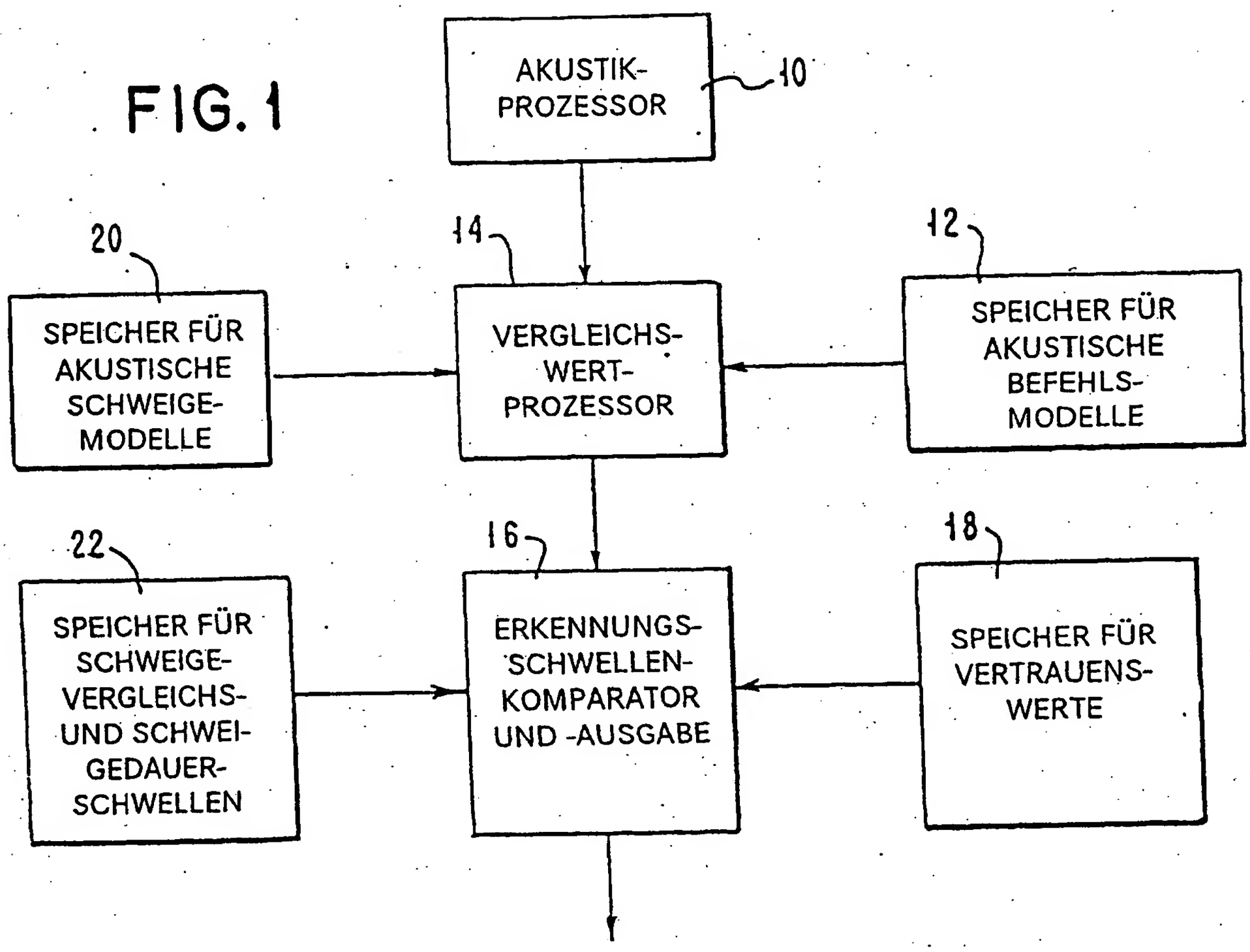
18. Spracherkennungsverfahren nach Anspruch 17, dadurch gekennzeichnet, dass die Anzeige für einen nicht erkennbaren Ton eines oder mehrere Fragezeichen umfasst.

19. Spracherkennungsverfahren nach Anspruch 11, dadurch gekennzeichnet, dass

jeder Ton einen Vokalton umfasst; und

jeder Befehl mindestens ein Wort umfasst.

FIG. 1


AKUSTIK-
PROZESSOR
10
20
SPEICHER FÜR
AKUSTISCHE
SCHWEIGE-
MODELLE
14
VERGLEICHS-
WERT-
PROZESSOR
12
SPEICHER FÜR
AKUSTISCHE
BEFEHLS-
MODELLE
22
SPEICHER FÜR
SCHWEIGE-
VERGLEICHS-
UND SCHWEI-
GEDAUER-
SCHWELLEN
16
ERKENNUNGS-
SCHWELLEN-
KOMPARATOR
UND -AUSGABE
18
SPEICHER FÜR
VERTRAUENS-
WERTE

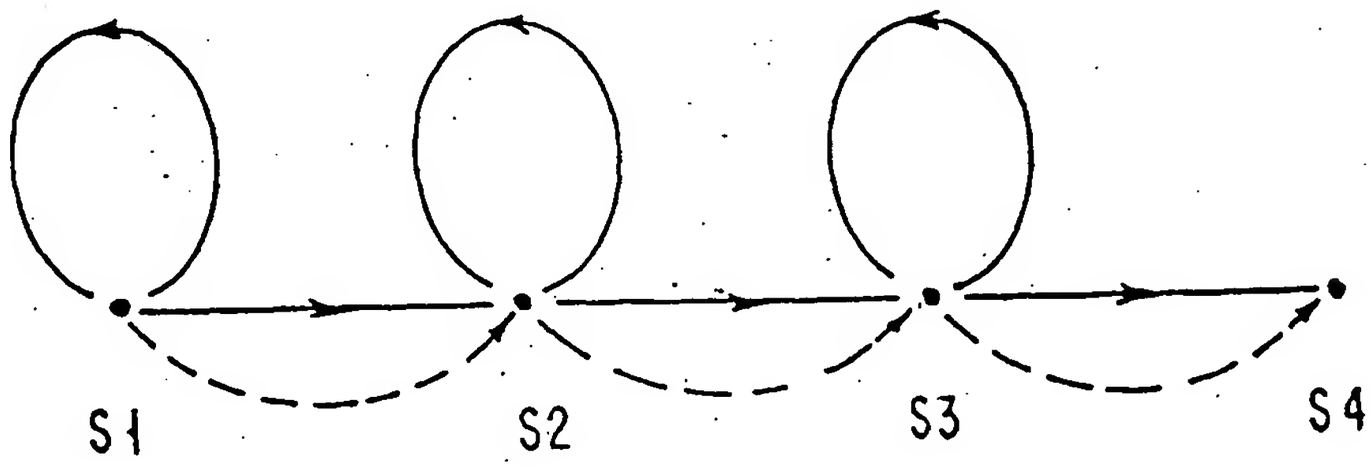

S1
S2
S3
S4

FIG. 2

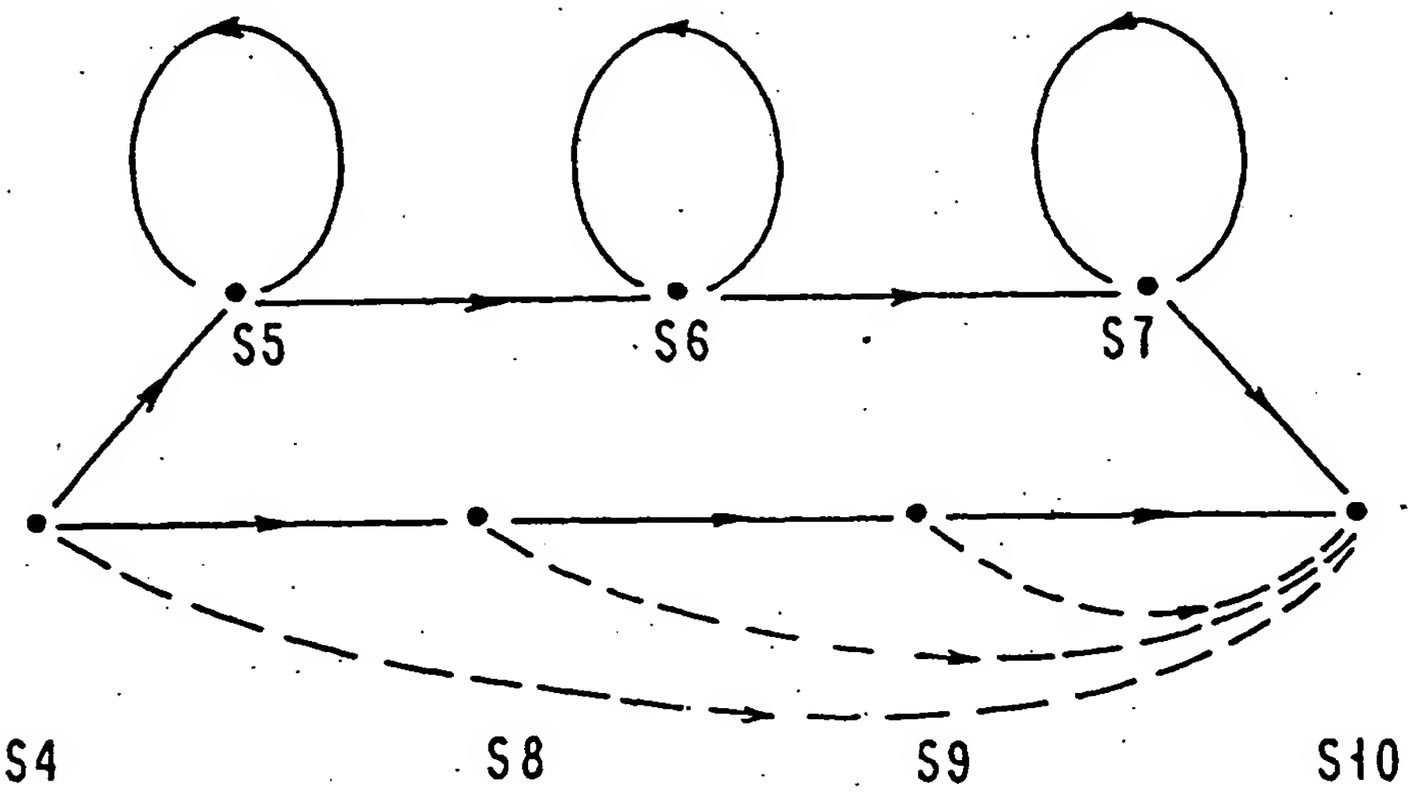


FIG. 3

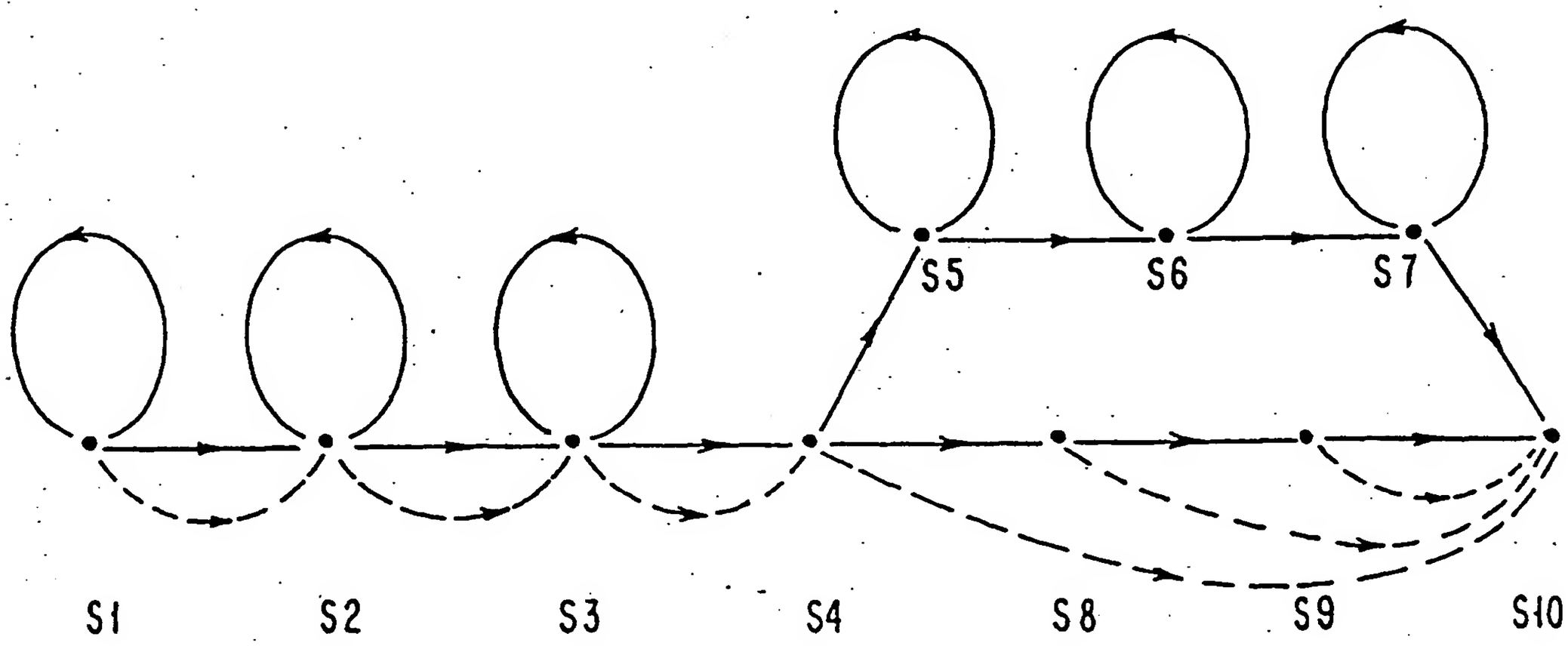


FIG. 4

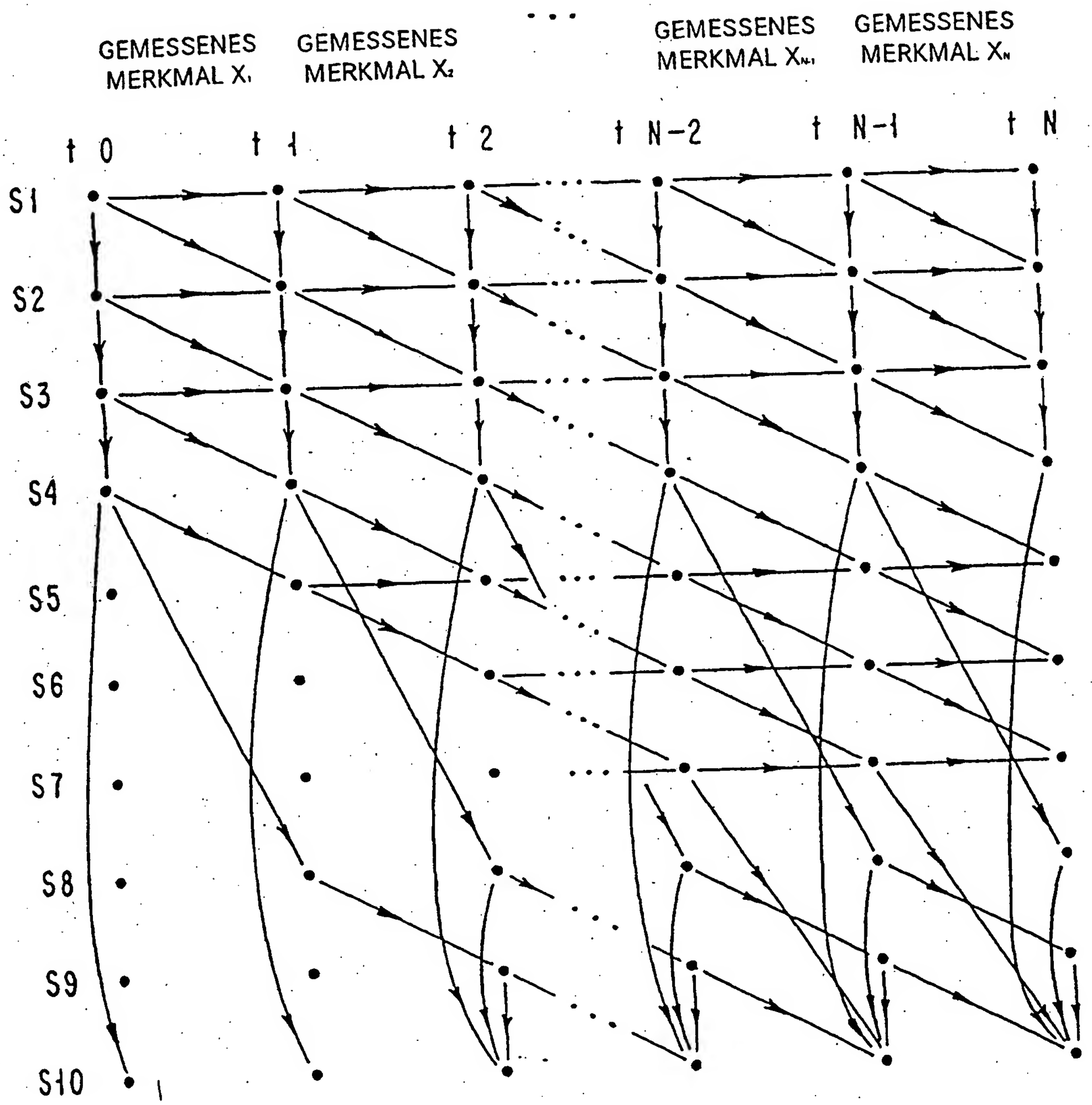

GEMESSENES MERKMAL X1
GEMESSENES MERKMAL X2
GEMESSENES MERKMAL XN-1
GEMESSENES MERKMAL XN
t 0
t 1
t 2
t N-2
t N-1
t N
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10

FIG. 5

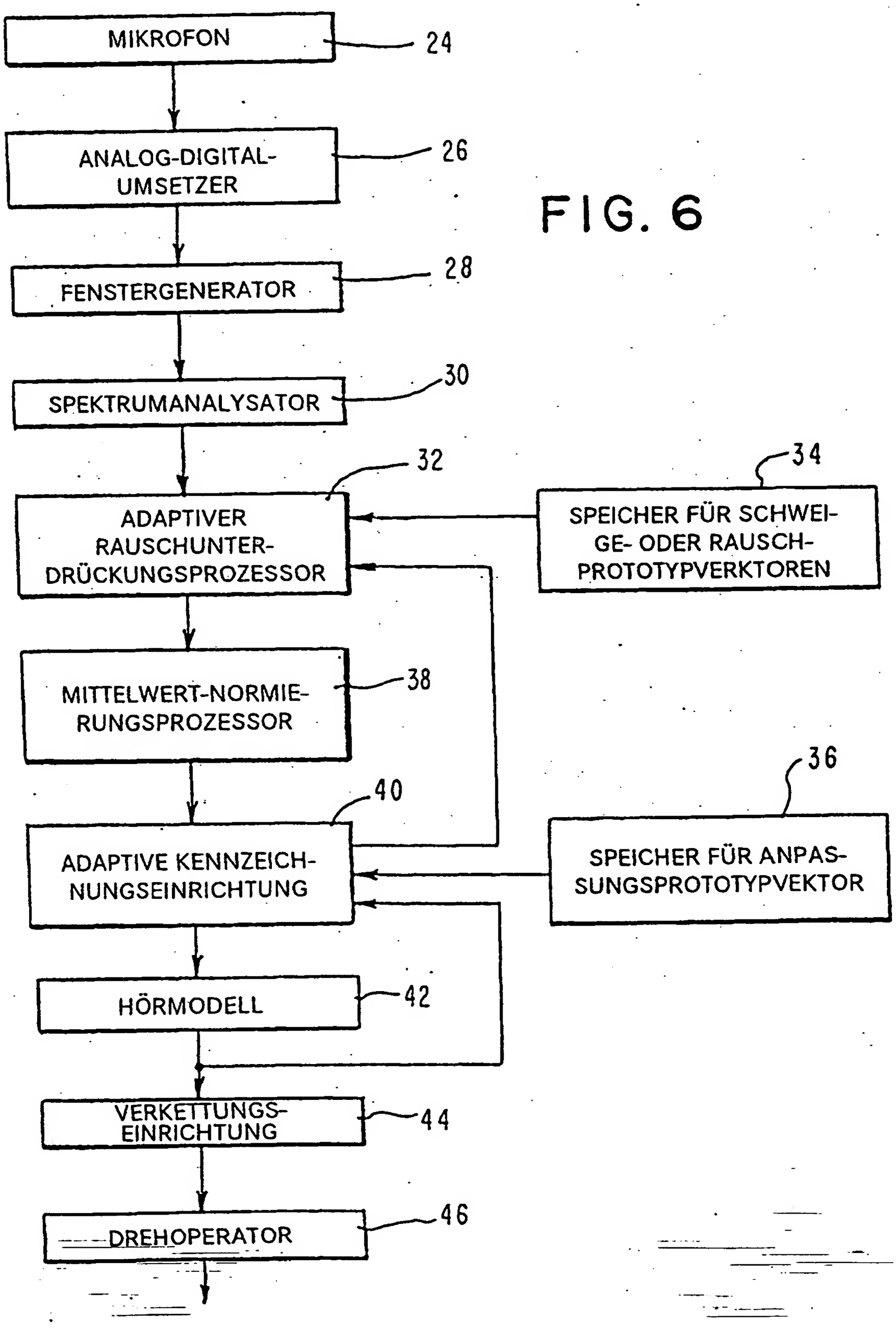

MIKROFON
24
ANALOG-DIGITAL-UMSETZER
26
FENSTERGENERATOR
28
SPEKTRUMANALYSATOR
30
ADAPTIVER RAUSCHUNTER-DRÜCKUNGSPROZESSOR
32
SPEICHER FÜR SCHWEI-GE- ODER RAUSCH-PROTOTYPVERKTOREN
34
MITTELWERT-NORMIE-RUNGSPROZESSOR
38
ADAPTIVE KENNZEICH-NUNGSEINRICHTUNG
40
SPEICHER FÜR ANPAS-SUNGSPROTOTYPVEKTOR
36
HÖRMODELL
42
VERKETTUNGS-EINRICHTUNG
44
DREHOPERATOR
46

FIG. 6

Docket # 2004P00824
Applic. #
Applicant: T. Engscheidt, et al.
Lerner Greenberg Stemer LLP
Post Office Box 2480
Hollywood, FL 33022-2480
Tel: (954) 925-1100 Fax: (954) 925-1101

Speech recognition system with improved rejection of words and sounds not contained in the system vocabulary.

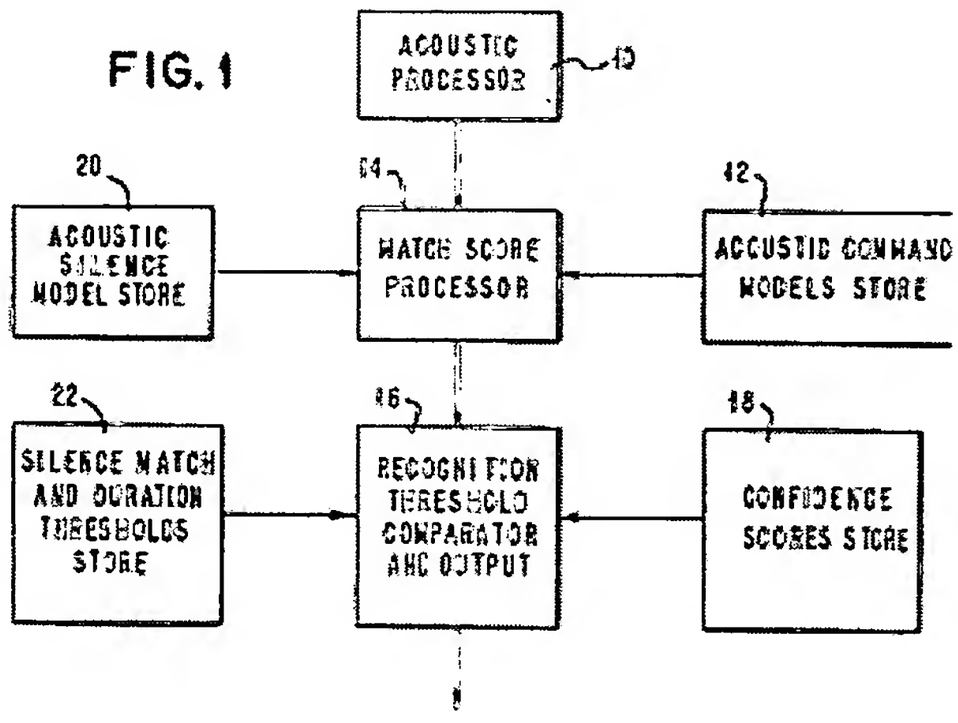
Publication number: DE69425776T
Publication date: 2001-04-12
Inventor: EPSTEIN EDWARD A (US)
Applicant: IBM (US)
Classification:
- international: ***G10L11/02;*** **G10L11/00;** (IPC1-7): G10L15/20
- european: G10L11/02
Application number: DE19946025776T 19940328
Priority number(s): US19930062972 19930518

Also published as:
EP0625775 (A
US5465317 (A
JP6332495 (A)
EP0625775 (B

Report a data error he

Abstract not available for DE69425776T
Abstract of corresponding document: **EP0625775**

A speech recognition apparatus and method output a recognition signal corresponding to a command model having the best match score for a current sound if the best match score for the current sound is better than a recognition threshold score for the current sound. The recognition threshold comprises a first confidence score if the best match score for a prior sound was better than a recognition threshold for that prior sound. The recognition threshold comprises a second confidence score better than the first confidence score if the best match score for a prior sound was worse than the recognition threshold for that prior sound. In one embodiment, the recognition threshold for the current sound comprises the first confidence score (a1) if the match score for the prior sound and the acoustic silence model is better than a silence match threshold, and if the prior sound has a duration exceeding a silence duration threshold, or (a2) if the match score for the prior sound and the acoustic silence model is better than the silence match threshold, and if the prior sound has a duration less than the silence duration threshold, and if the best match score for the next prior sound and an acoustic command model was better than a recognition threshold for that next prior sound, or (a3) if the match score for the prior sound and the acoustic silence model is worse than the silence match threshold, and if the best match score for the prior sound and an acoustic command model was better than a recognition threshold for that prior sound. The recognition threshold for the current sound comprises the second confidence score better than the first confidence score (b1) if the match score for the prior sound and the acoustic silence model is better than the silence match threshold, and if the prior sound has a duration less than the silence duration threshold, and if the best match score for the next prior



sound and an acoustic command model was worse than the recognition threshold for that next prior sound, or (b2) if the match score for the prior sound and the acoustic silence model is worse than the silence match threshold, and if the best match score for the prior sound and an acoustic command model was worse than the recognition threshold for that prior sound.

Docket # 2004P00324

Applic. #

Applicant: T. Fingscheidt, et al.

Lerner Greenberg Stemer LLP
Post Office Box 2480
Hollywood, FL 33022-2480
Tel: (954) 925-1100 Fax: (954) 925-1101